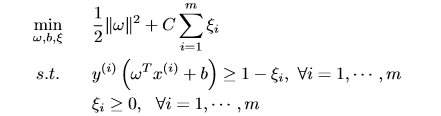
**实验五：SVM**

1、描述

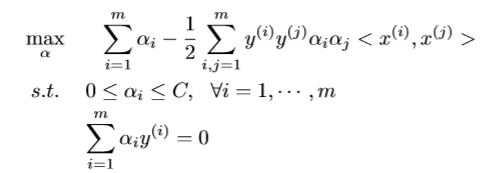
这个练习让你练习使用SVM进行线性和非线性分类。

2、SVM

第一部分是实现正则化的SVM分类算法。它的细节可以在lecture slides中找到。下面我们只给出一个关于支持向量机的草图。正则化SVM可以表述为：



其中 ξi 是松弛变量。它的双重性问题可以归结为：



通过求解上述QP问题，可以得到一个SVM分类器。为了简单起见，我们可以使用现有的QP求解器（例如，MATLAB中的四叉树或Octave中的qp）。

给出了两个数据集，每个数据集分别分为训练和测试两部分。特别地，rst数据集包括用于训练的traing 1.txt和用于测试的test 1.txt，而第二数据集包括traing 2.txt和test 2.txt。在这些数据中，rst两列是特征，最后一列是label.分别在两个数据集上尝试SVM，并回答以下问题：  
1、用训练数据绘制SVM的决策边界。

1. 使用测试数据评估SVM分类器，并显示被错误分类的测试示例部分。
2. 尝试正则化项C的二元数值，并报告您的观察结果。

3、手写数字识别

第二部分是应用SVM分类器识别手写数字。在 train-01-image.svm 和 test01-images.svm中给出了用于训练和测试的数据集。为了简单起见，我们只区分0和1。训练数据集包括12665幅图像，而用于测试的数据集包括2115幅图像。每行表示一个图像，其中rst项是标签，而以下是像素的索引和对应的灰度值。注意，仅给出具有非零灰度值的像素。请仔细阅读并尝试strimage.m，其中给出了如何处理数据示例。在报表中显示一些数据示例。根据训练数据训练SVM模型，并将其应用于识别测试数据集中给出的手写数字。回答以下问题：

1. 在train-01-images.svm上训练一个普通的SVM（即，对于∀i，ξi=0没有正则化）。什么是训练误差，即训练数据中被错误分类的示例部分。从一组错误分类图像中挑选一个。 附上它的plot与你的解决方案。为什么你认为SVM在训练过程中没有正确分类？现在将SVM应用于测试集test-01-image.svm；测试错误是什么，即错误分类的测试数据的部分？
2. 试验正则化项C的二元数值。从猜测/估计你认为C应该存在的范围开始。然后选择C值（在该范围内），您将在该值上评估SVM的性能。您不需要选择超过10个这样的值，尽管您应该可以随意选择这么多（或更少！）如你所愿。对于C的这些值，绘制相应的误差。
3. C的哪一个值给您从部分数据集上的最佳训练错误？
4. 这种选择C的测试误差与您在(i)部分计算的测试误差相比如何？

（iii）你能优化C使其产生最好的测试错误而不是训练错误吗？你应该吗？

4、非线性SVM

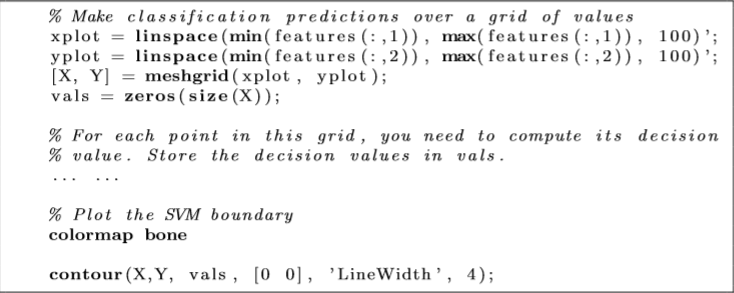
实现非线性SVM的一种方法是引入内核。回想一下，线性不可分离特征在映射到高维特征空间之后常常变得线性可分离。然而，我们永远不需要显式地计算特征映射φ(x(i))：我们只需要处理它们的内核，它们更容易计算。因此，可以基于高维（甚至内维）特征映射来创建非常复杂的决策边界，但是由于内核表示，仍然具有复杂的计算。

在本练习中，您将把径向基函数（RBF）内核（而不是正则化方法）应用到SVM模型。这个内核有公式：  
 

注意，这与高斯内核是一样的，只是高斯核中的项已经被γ代替了。再一次，请记住，在任何时候，您都不需要直接计算φ(x)。实际上，φ(x)对于这个内核来说是无穷维的，所以不可能将它存储在内存中。

现在让我们看看RBF核如何选择一个非线性决策边界。将数据集训练 training 3.txt加载到Matlab/OcthVE工作空间中。这是一个二维分类问题。用不同的颜色绘制正反两种颜色。这个数据集是否存在线性决策边界？

根据上述练习，在γ=100的RBF核上训练SVM模型。一旦有了模型，您就需要可视化决策边界，参考以下代码：



回想一下，函数给出用于进行分类的决策值。如果，示例x被分类为正示例，否则被分类为负。现在使用1、10、100和1000的γ值训练模型，并为每个模型绘制决策边界（不使用等高线ll）。边界的T如何随γ变化？

5、序贯最小优化

这是一个可选的练习。如果可以使用序列最小优化（Se.ialMinimal.，SMO）算法来替换标准QP求解器，则可以获得额外的学分。